



# A roadmap to Durable BCTV Resistance Using Long-Read Genome Assembly of Genetic Stock KDH13

Paul J. Galewski<sup>1</sup> · Imad Eujayl<sup>1</sup>

Received: 28 September 2020 / Accepted: 18 July 2021

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2021

## Abstract

Beet Curly Top (BCT) is a viral disease which negatively impacts crop productivity for sugar beet growers and the sugar beet industry in the western USA and dry regions worldwide. Current varieties exhibit little genetic resistance to the Beet Curly Top Virus (BCTV), suggesting there is a large potential for improvement. KDH13 (PI 663862) is a double-haploid line created from a population (C762-17/PI 560130) which segregates for resistance to BCTV and was identified as genetic stock for the improvement of sugar beet varieties. PacBio sequences were generated and assembled to better define the content and organization of variation within the KDH13 genome and to provide resources to identify specific variation underpinning durable genetic resistance. Using ab initio predicted proteins as anchors, the assembled KDH13 contigs were placed in a more contiguous order using the EL10.1 reference genome, which leveraged Bio-Nano optical maps and Hi-C proximity information for chromosome level scaffolding. In total, 4681 (75%) of the 6245 contigs were placed in the order and orientation of the EL10.1 genome. The anchored contigs represented 502,929,268 bp (87.7%), the KDH13 genome assembly. An F<sub>1</sub> hybrid and parental lines KDH13 (resistant) and KDH19-17 (susceptible) were sequenced using Illumina technology in order to characterize the SNP, indel, and structural variation between parental lines and allow for a more detailed investigation into causal variation linked to important phenotypes. In total, 3,086,720 variants were detected, including 2,259,324 single-nucleotide polymorphisms, 191,448 insertions, 198,057 deletions, 268,090 complex substitutions, 90,004 multi allelic variants, and 79,797 structural variants. Of the total variation, 1,158,491 were informative in the F<sub>1</sub> and were able to discriminate between the two parents. This information represents a high-density marker dataset distributed globally across the sugar beet genome and can be used to track genomic segments in populations where KDH13 is used as parental material to improve BCTV resistance.

**Keywords** Sugar Beet · Genome · PacBio · Curly top · Disease resistance

## Key Message

- Developing crop varieties with durable resistance to current and emerging pathogens is a critical challenge for maintaining agricultural productivity
- The development of useful breeding materials is an arduous and time-consuming task
- Genomics has turbocharged these efforts, and well-developed gene ontologies have afforded a better perspective on gene function and molecular mechanisms underpinning important agronomic characters
- Modern approaches to plant breeding have the potential to address long-standing problems with agricultural productivity and sustainability but requires sequence data from biological samples which contain characters with a high potential for addressing these problems

✉ Paul J. Galewski  
paul.galewski@usda.gov

<sup>1</sup> USDA-ARS - Northwest Irrigation and Soils Research Laboratory, Kimberly, ID, USA

## Introduction

Beet Curly Top (BCT) is a viral disease prevalent throughout the Western US and Middle East. The beet curly top virus (BCTV) is a parthenogenic Curtovirus transmitted by the beet leafhopper (*Circulifer tenellus*). BCTV infection can cause devastating losses for beet growers and threatens the profitability of the sugar industry in areas where the disease is common. Currently, BCT disease pressure is managed by a combination of genetic resistance present at varying levels within commercial varieties and management practices in order to control vector populations such as insecticide sprays and seed treatments (e.g., Poncho-Beta, Cruiser and NipsIT) (Strausbaugh et al. 2012, 2014). Genetic resistance to BCTV is a quantitative trait (Panella et al. 2014), and the breeding and maintenance of varieties with durable genetic resistance is complex due to the need to continuously incorporate resistance genes for multiple pathogens simultaneously. Breeding varieties with durable genetic resistance is becoming more important due to the potential loss of

effective chemicals to control the disease in fields which will likely increase in severity and as a result could compromise the productivity and profitability of the crop for growers (Strausbaugh et al. 2006). A fundamental understanding of the molecular mechanisms (e.g., genome variation, genes, and pathways) could allow for the development of a genome informed approach to improve resistance in varieties and address this long-standing challenge to sugar beet cultivation in regions where BCTV is problematic.

The current state of genomic resource development for beet is ongoing. A plethora of research in diverse systems has demonstrated how this technology can interface directly with crop improvement (Bevan et al. 2017). The construction of reference genomes is also evolving; long-read technologies provide solutions to the challenges encountered when assembling highly repetitive plant genomes using short reads. Several reference genomes exist for beet which highlights this progress (Dohm et al. 2014; McGrath et al. 2020). High-quality reference genomes in beet have provided a lens to measure genome diversity (Galewski and McGrath 2020) as well as assess the content of important functional genes such as those involved in disease resistance (Funk et al. 2018). Comparing genome assemblies requires relatively complete genome sequences in order to evaluate the content and organization of genomic variation within genomes. Only recently has sequence data become available which looks at genome diversity in primary (Galewski and McGrath 2020), secondary (Galewski and McGrath 2020), and tertiary (del Rio et al. 2019) gene pools of sugar beet.

PacBio long reads have resulted in more contiguous assemblies and provides a means to investigate larger structural variation, inversions, and their resulting phenotypic consequences. Structural variation (SV), including indels, likely contributes more to lineage divergence than single-nucleotide polymorphisms (SNPs) due to a larger potential for these variants to alter gene function (Chen et al. 2009). Long reads have a greater ability to detect SV and chromosomal inversions than short reads. Recent studies have shown associations between large chromosome-scale structural variants with adaptive and agronomic trait variation in plants (Schiessl et al. 2019). In *Brassica napus*, up to 10% of all genes were affected SV events. The complex nature and highly duplicated nature of plant genomes makes long-read technology well suited for addressing SV in a high-throughput manner (Chawla et al. 2020).

One hope of developing genomic resources in beet is that useful genetic markers can be extracted to facilitate in the breeding and improvement process. Genetic markers can facilitate the introgression of important characters such as BCTV resistance from resource populations (e.g., mapping populations, germplasm resources, and wild relatives) into elite cultivars. Generating sequence data that captures causal variation should be considered in any project which can lead to a better understanding of the underlying molecular mechanisms, provide targets for gene manipulation, and validation of gene function. This

paper represents a pipeline for the fast and efficient characterization of beet genomes using Pac-bio long-read sequencing to ultimately inform the genetic basis of important agronomic characters. The double-haploid KDH13 (PI 663862) was created from a single plant selected for BCTV resistance from the population C762-17 (PI 560130) (Eujayl et al. 2016). This line was sequenced and assembled to capture the genetic variation associated with BCTV resistance. Additionally, Illumina sequencing data was generated for KDH13, KDH19-17, and an F<sub>1</sub> hybrid (KDH13×KDH19-17) in order to determine informative genomic markers that can be used in the analysis of F<sub>2</sub> families and breeding for improved BCTV resistance.

## Materials and Methods

### KDH13 Plant Material

Seeds from the sugar beet (*Beta vulgaris* L.) doubled-haploid line KDH13 (PI 663,862) were planted in Sungro Horticulture Professional Growing Mix, containing Canadian sphagnum peat moss, perlite, dolomite lime, and a long-lasting wetting agent, RESILIENCE. Beets were grown in a growth control chamber at 22 °C, then moved to a greenhouse after three and a half months.

### KDH13 Genome Sequencing

A single KDH13 sugar beet plant was dark treated for 48 h in order to enhance the efficiency of DNA extraction by depleting sugar reserves and inhibiting secondary metabolite synthesis. Whole leaves were placed into 50-ml centrifuge tubes and filled with liquid nitrogen and stored at −80°. Samples were shipped overnight on dry ice to GENEWIZ Plainfield (NJ, USA) for nuclei isolation, extraction of high molecular weight (HMW) DNA, and library preparation. A 20-kb SMRTBell Library was prepared and sequenced using a total of 3 SMRTcells on the PacBio Sequel platform. Read lengths generated from the PacBio sequencing run were plotted using the program QUAST (Mikheenko et al. 2018).

### KDH13 Genome Assembly

The resulting reads were used to assemble a de novo genome for the BCTV resistant double-haploid, KDH13, using the program Canu1.7 (Jayakumar and Sakakibara 2019). Canu1.7 was run with default parameters using an expected genome size of 758 Mb (Arumuganathan and Earle 1991).

### KDH13 Genome Evaluation

Several programs were used to evaluate the KDH13 assembly and compare with other current beet genome assemblies

(e.g., RefBeet1.2 and EL10.1), representing genome sequences resulting from different genetic backgrounds and sequencing and technologies. Genome statistics were generated using the script `assemblathon.pl` (<https://github.com/ucdavis-bioinformatics/assemblathon2-analysis>) (Bradnam et al. 2013). The program BUSCO was used to evaluate the completeness of these genomes by the identification of kingdom-specific core orthologs found across other “gold standard” plant reference genomes (Seppey et al. 2019).

### KDH13 Functional Genome Predictions

Ab initio gene prediction was carried out on the KDH13 genome assembly using AUGUSTUS version 3.3.2. Current HMM models representing the Pfam-A dataset were downloaded from <ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam33.1/Pfam-A.hmm.gz>, and the predicted proteins were annotated using the program PfamScan which used the HMM protein family signatures and domains within the KDH13 genes predicted with AUGUSTUS. The program liftOff (Shumate and Salzberg 2020) provided additional evidence to gene model structures by transferring the annotations from the reference EL10.1 to the KDH13 genome assembly.

### Genome Order

The program Blastx was used to identify physical positions of the protein predictions within the EL10.1 genome. Positional information in the resulting blast file allowed the proteins to be used as anchors to determine a more contiguous order of KDH13 contigs to the EL10.1 genome, which was scaffolded using proximity information such as Hi-C and BioNano optical maps. Custom scripts were used to evaluate the output of Blastx using the KDH13 predicted proteins with the ordered EL10.1 protein prediction and determine a more contiguous order of the KDH13 assembly, which related to accepted linkage groups (Schondelmaier and Jung 1997) and the chromosome nomenclature detailed in Butterfass 1964. Additionally, the ordered genome produced was validated by comparing the order of the EL10.1 genome annotation transferred using the program liftOff (Shumate and Salzberg 2020).

### Analysis of Genome Variation

This variation was used to assess unique variation within KDH13 genome. Illumina genome sequencing was carried out for three different genotypes, KDH13, KDH19-17, and F<sub>1</sub> (KDH13 × KDH19-17). All reads were filtered and trimmed in a sliding window based on the average

quality parameter ( $Q > 20$ ) using Trimmomatic (Bolger et al. 2014). Filtered reads were used for all downstream analysis. BWA (Li 2013) was used to align Illumina short-read data for DH line KDH13 to the EL10.1 genome. Variant calling was performed on the bam files using the program Freebayes (Garrison and Marth 2012). Variants were filtered for quality and read depth (e.g., read depth,  $N > 10$  and  $N < 300$ ,  $QUAL \geq 20$ , and mapping quality  $\geq 60$ ). The program Manta (Chen et al. 2016) was used to detect structural variation within KDH13, KDH19-17, and F<sub>1</sub> (KDH13 × KDH19-17) using Illumina data. The density of variation between the two parents KDH13 and KDH19-17 was estimated to give an idea of the divergence between the two parental lines. Genomescope 2.0 (Ranallo-Benavidez et al. 2020) was used to evaluate k-mer spectra produced by the program Jellyfish (Marçais and Kingsford 2011) for each Illumina dataset. The K-mer data provided estimates of the size of the genomes, content of repetitive sequences, and the quantity and distribution of heterozygosity.

### Relationship Determination

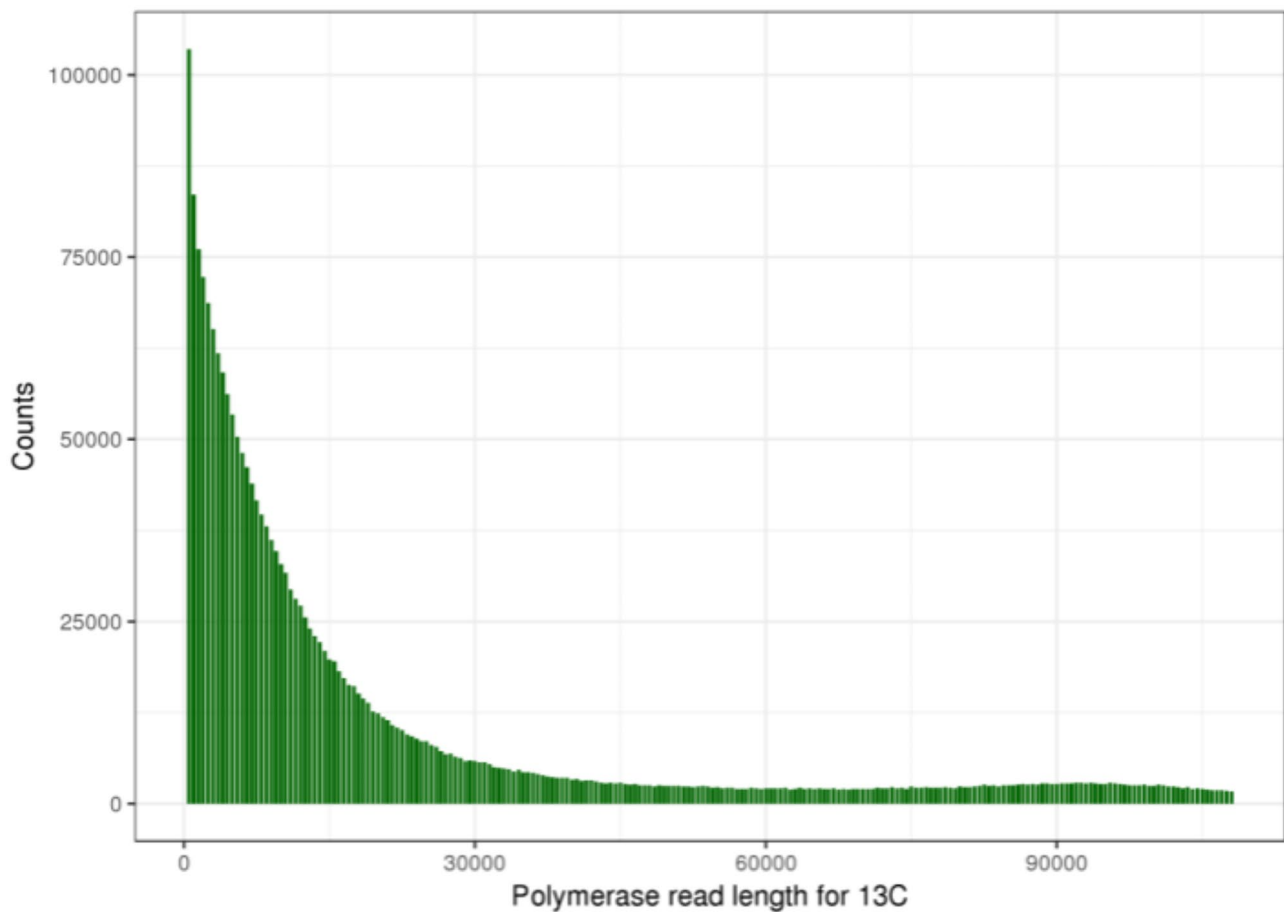
Biallelic SNPs were used to calculate pairwise relationship coefficients between lineages (both inbred and outbred) using the Kinship Inference for Association Genetic Studies (KING) package (Manichaikul et al. 2010). Data for additional beet populations also used in clustering was derived from Galewski and McGrath (2020).

### Visualization of KDH13 Genome (Python and R)

Genome order was deduced from alignment of KDH13 predicted proteins with known locations on assembly contigs with scaffolds of the EL10.1 genome assembly, effectively using the protein sequence as anchors and custom scripts to determine the position and order of predicted proteins within the EL10.1 genome. Custom R scripts were written to visualize the KDH13 genome relative to other beet genomes (e.g. *Beta vulgaris* crop types (Galewski and McGrath 2020), F<sub>1</sub> (KDH13 × KDH19-17), and KDH19-17 lineages. Gene density was extracted from the EL10.1 gene set. Gypsy and Copia repeat density was extracted from LRT-retriever (Ou and Jiang 2018). The centromeric repeat Beetle7 was used in cytogenetic studies of beet (Kowar et al. 2016) and was located within the EL10.1 genome using Blastn. The blast results were parsed and plotted against the EL10.1 genome.

### Phenotypic Data

Phenotypic evaluation of curly top resistance was carried out using the scale detailed in Mumford 1974. Individual



**Fig. 1** Read length distribution of PacBio sequence data

beets from the resistant parent (KDH13) susceptible parent (KDH19-17) and  $F_1$  populations were grown in the greenhouse, infected with BCTV via beet leaf hoppers and evaluated for disease severity.

## Results

PacBio sequencing was used to generate long reads (~20 kb) for the sugar beet double-haploid KDH13. Using these reads, a genome was assembled using the program Canu 1.7. This new reference genome assembly served to capture sequence

variation for a genome associated with resistance to the BCTV. To evaluate the quality of the KDH13 genome, comparisons with other published beet genomes were carried out against the EL10.1 and RefBeet-1.2.2 genomes. The KDH13 contigs were ordered relative to the EL10.1 assembly using predicted proteins as anchors to find their genomic locations within EL10.1. Finally, Illumina data was used to give insight into the content and distribution of genome variation contained within genomes of KDH13, KDH19-17, and an  $F_1$  hybrid (KDH13  $\times$  KDH19-17). These data produced informative markers for characterizing these lines and for evaluating the segregation of durable genetic resistance to BCTV in  $F_2$ -derived populations.

**Table 1** Description of the previously sequenced *Beta vulgaris* genomes and materials

Assembly	Plant material	Accession	Sequencing technology	Source
EL10.1	Self-fertile inbred	C869 (PI 628755)	Illumina, PacBio, Bio-Nano, Hi-C	McGrath et al. (2020)
KDH13	Double haploid	KDH13 (PI 560130)	PacBio	-
RefBeet-1.2.2	Double haploid	KWS2320	Illumina short read technology	Dohm et al. (2014)

**Table 2** *Beta vulgaris* genome assembly comparisons

Assembly	EL10.1	KDH13	RefBeet-1.2.2
Number of scaffolds	9	6246	40,246
Total size of scaffolds	520,115,771	573,479,205	566,550,431
Longest scaffold	65,096,967	1,040,497	60,962,716
Shortest scaffold	52,180,088	1009	499
Number of scaffolds > 1 K nt	9(100%)	6246(100%)	8154(20.3%)
Number of scaffolds > 10 K nt	9(100%)	6013(96.3%)	868(2.2%)
Number of scaffolds > 100 K nt	9(100%)	1893(30.3%)	300(0.7%)
Number of scaffolds > 1 M nt	9(100%)	1893(0%)	38(0.1%)
Number of scaffolds > 10 M nt	9(100%)	1893(0%)	9(0%)
Mean scaffold size	57,790,641	91,815	14,077
Median scaffold size	57,938,902	53,481	727
N50 scaffold length	57,938,902	163,423	34,941,034
L50 scaffold count	5	1032	7
Scaffold %A	32.07	32.14	29.16
Scaffold %C	17.92	17.89	16.51
Scaffold %G	17.91	17.85	16.53
Scaffold %T	32.09	32.12	29.2
Scaffold %N	0.01	0	8.61
Scaffold %non-ACGTN	0	0	0
Number of scaffold non-ACGTN nt	0	0	0
Percentage of assembly in scaffolded contigs	100.00%	0.00%	93.60%
Percentage of assembly in unscaffolded contigs	0.00%	100.00%	6.40%
Average number of contigs per scaffold	37	1	1.5
Average length of break (> 25 Ns) between contigs in scaffold	178	0	2455
Number of contigs	333	6246	60,090
Number of contigs in scaffolds	333	0	20,872
Number of contigs not in scaffolds	0	6246	39,218
Total size of contigs	520,057,958	573,479,205	517,832,498
Longest contig	13,084,595	1,040,497	449,869
Shortest contig	0	1009	154
Number of contigs > 1 K nt	330	6246	27,257(45.4%)
Number of contigs > 10 K nt	330	6013	11,611(19.3%)
Number of contigs > 100 K nt	325	1893	574(1%)
Number of contigs > 1 M nt	162	1	0(0%)
Number of contigs > 10 M nt	1	0	0(0%)
Mean contig size	1,561,736	91,815	8618
Median contig size	974,632	53,481	918
N50 contig length	2,772,180	163,423	43,432
L50 contig count	60	1032	3471
Contig %A	32.07	32.14	31.9
Contig %C	17.92	17.89	18.06
Contig %G	17.91	17.85	18.08
Contig %T	32.09	32.12	31.95
Contig %N	0	0	0.01
Contig %non-ACGTN	0	0	0
Number of contig non-ACGTN nt	0	0	0

## KDH13 Sequencing, Genome Assembly, and Evaluation

In total, 14.8 Gb was produced from the PacBio Sequel II sequencing platform resulting in an average genome wide

coverage of 19.54× using an estimated genome size of 758 Mb. The average read length was 20,901 bp (Fig. 1). The de novo genome assembled from PacBio reads resulted in a genome size of 573,479,205 bp in length. The Assemblethon.pl script allowed direct comparison of genome assemblies between



**Table 3** Results of BUSCO analysis

	EL10.1	RefBeet-1.2.2	KDH13
<b>Genome size</b>	520,115,771	566,550,431	573,479,205
<b>Complete BUSCOs (C)</b>	413(97.2%)	421(99.1%)	360(84.7%)
<b>Complete and single-copy BUSCOs (S)</b>	407(95.8%)	413(97.2%)	354(83.3%)
<b>Complete and duplicated BUSCOs (D)</b>	6(1.4%)	8(1.9%)	6(1.4%)
<b>Fragmented BUSCOs (F)</b>	1(0.2%)	3(0.7%)	34(8.0%)
<b>Missing BUSCOs (M)</b>	11(2.6%)	1(0.2%)	31(7.3%)
<b>Total BUSCO groups searched</b>	425	425	425

different genotypes and sequencing strategies (Table 1). Comparisons between KDH13 and previously sequenced reference genomes show similar genome sizes were recovered by all methods (520–573 Mb) (Table 2). Consistently, these values fall short of the 758 Mb predicted size. Long-read single-molecule sequencing technology shows a large reduction in the number of scaffolds in PacBio KDH13 genome assembly ( $N=6246$ ) versus Illumina RefBeet-1.2.2 (40,246). The EL10.1 genome which in addition to PacBio reads used Bio Nano optical maps and Hi-C proximity information, produced a scaffold number equal to the haploid chromosome number ( $N=9$ ). Similarities in nucleotide percentages were observed between genomes with a reduction in RefBeet-1.2.2 as a result of the addition of Ns to the genome in order to link scaffolds into pseudomolecules using a linkage map.

### Gene Content Comparisons

BUSCO compared the presence of 425 conserved ortholog genes across the different assemblies. The presence of core ortholog genes tests the completeness of each genome assembly. This may result from differences in assembly methods and sequencing strategies rather than genotypes used. RefBeet-1.2.2 showed the largest number of complete BUSCOs, followed by EL10.1. The KDH13 genome was less complete than EL10.1 or RefBeet-1.2.2 and had the greatest proportion of missing BUSCOs (7.3%) and fragmented BUSCOs (8.0%). The KDH13 genome did not use Illumina short-read data in the assembly (Table 3).

### Functional Genome Predictions

Ab initio gene prediction using AUGUSTUS produced 80,342 protein coding genes. Based on what has been recovered in EL10.1 (24,255) and RefBeet-1.2.2 (27,421), this figure is high, suggesting additional CDNA and EST evidence for accurate genome annotation. Of these predicted genes, 40,779 contained one or more Pfam domains and a blast of these proteins to the EL10.1 predicted protein set showed 27,847 of the predicted

proteins aligned to the EL10.1 protein set with confidence levels assessed by both e-values  $< 1 \times 10^{-60}$  and a sequence similarity  $> 75\%$ . Additionally, gene annotations were transferred from the EL10.1 genome to KDH13 using the program liftOff. In total, 23,389 (96.4%) of the 24,255 genes from the EL10.1 were placed on the KDH13 genome assembly. Together, these datasets offered a picture of the gene content, protein coding capacity, and functional regions within the KDH13 genome.

### Order of Genome Contigs

The protein predictions from the KDH13 genome were used as anchors to place KDH13 contigs in the order and orientation of the EL10.1 genome. For contigs with one or more protein predictions, blastx results were evaluated for percent identity and e-value to determine a genomic location within EL10.1. In total, 4681 of the 6245 contigs (75%) were placed in the order and orientation of the EL10.1 genome which was determined using Bio Nano optical maps and Hi-C proximity information. The number of contigs placed on each chromosome is present in Table 4. The 25% of contigs which showed less confidence for placement in a more contiguous order were left as unscaffolded. The contigs that were placed in the correct sequence were equal to 502,929,268 bp or 87.7% of the total size of the KDH13 genome assembly. The success of this approach was

**Table 4** Number of contigs placed in the order and position relative to the EL10.1 reference genome

Chromosome	Number of contigs
Chromosome 1	404
Chromosome 2	396
Chromosome 3	437
Chromosome 4	528
Chromosome 5	425
Chromosome 6	455
Chromosome 7	403
Chromosome 8	371
Chromosome 9	456
Scaffolds (10–22)	504
Total	4681

**Table 5** Alignment of KDH13 Illumina sequence data to EL10.1 genome using different quality thresholds

Filter	N (bp)	% coverage
Map Q > 60	471,933,082	0.907
Map Q > 20	480,543,985	0.924
Map Q > 0	492,161,490	0.946
No filter	520,115,771	

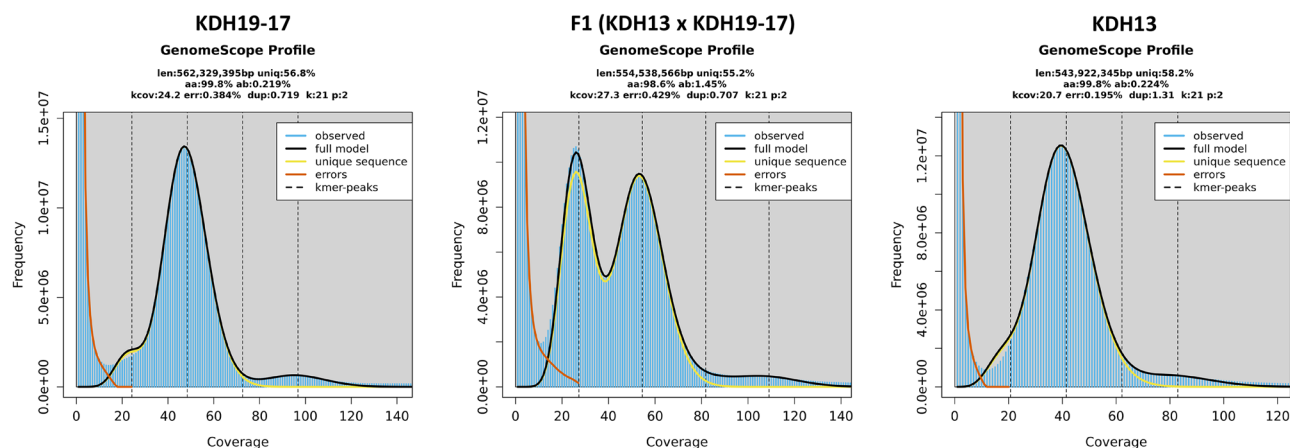
confirmed by the congruence with the order of EL10.1 gene order produced by liftoff. The unanchored contigs were difficult to place in the genome using either method.

## Genome Variation

KDH13 Illumina reads were mapped to the EL10.1 genome. As quality thresholds were adjusted, coverage was reduced, suggesting up to 90.7% was highly similar and 5.4% of the genome had no coverage even using low mapping quality thresholds. This small percentage may account for the divergence between KDH13 and EL10.1 as alignments were not predicted due to the accumulation of sequence polymorphism but may result from sequencing error (Table 5). Mean depth across the genome was 22.34× of the EL10.1 genome. The standard deviation was 122.91 with a depth of coverage ranging from 1 to 9061 across sites. Genomescope 2.0 showed the k-mer spectra of the F<sub>1</sub> (KDH13×KDH19-17) was intermediate to the respective parents (Fig. 2). The genome size, degree of homozygosity, and number of repetitive sequences are detailed in Table 6. The heterozygosity of the F<sub>1</sub> was 1.43%. This represents a sevenfold increase in heterozygous loci relative to the double-haploid parents KDH13 and KDH19-17 (0.19% and 0.20%).

## KDH13, KDH19-17, and F<sub>1</sub>(KDH13×KDH19-17) Genome Variation

A variant calling and filtering pipeline was used to determine the variation specific to KDH13 and KDH19-17 populations. Unique variation contained within the KDH13 genome likely conditions or is linked to BCTV resistance. In total, 3,086,720 variants were detected. Of these, 2,259,324 (73.2%) were classified as SNP, 191,448 (6.2%) as insertions, 198,057 (6.4%) as deletions, 268,090 (8.7%) as complex substitutions, 90,004 (2.9%) as multi nucleotide polymorphisms (MNP), and 79,797 (2.6%) as SV (Table 7). Of the total variants detected in the parental genomes KDH13 and KDH19-17, 76,388 (6.1%) and 234,262 (13.2%) were detected as heterozygous respectively. Several regions appeared to contain greater heterozygosity than expected across the genome. Most notably were regions on chromosome 3 and chromosome 7 for KDH19-17 (Fig. 3). The variants were plotted to determine if this result was biological (e.g., repetitive regions) or technical (e.g., assembly or alignment) in nature. The similarity between the KDH13 and EL10.1 genome was apparent. Less variants were detected in KDH13 relative to KDH19-17. The number of variants that can discriminate between KDH13 and KDH19-17 was estimated to be 1,272,863 (40.2%). The additional variant sites were similar between two parents and lacked the ability to discriminate between parents either as a result of fixation for the same variant allele and/or segregation within the parental populations. Relationships were determined between KDH13, KDH19-17, F<sub>1</sub> and various accessions representing cultivated *Beta vulgaris* (Fig. 4). A major difference in hierarchical clustering suggests breeding history plays a major part in divergence. Two groups could be identified: an inbreed group consisting of C869, W357B, KDH13, KDH19-17, F<sub>1</sub> and an outbreed group of pooled populations. Divergence via inbreeding appeared to have a large effect on clustering, which appears to be more important than shared

**Fig. 2** K-mer spectra of KDH19-17, F<sub>1</sub> (KDH13×KDH19-17), and KDH13

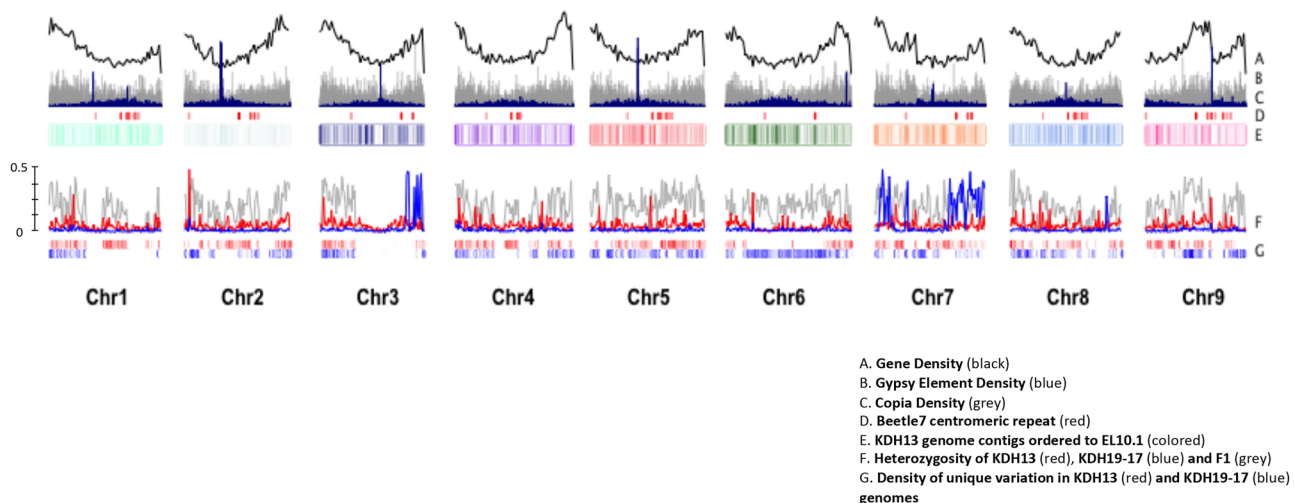
**Table 6** Results of K-mer spectra analysis using Jellyfish and Genomescope 2.0

	KDH19-17		KDH13 × KDH19-17		KDH13	
	Min	Max	Min	Max	Min	Max
<b>Homozygous (aa)</b>	99.76%	99.81%	98.54%	98.57%	99.75%	99.80%
<b>Heterozygous (ab)</b>	0.19%	0.24%	1.43%	1.46%	0.20%	0.25%
<b>Genome haploid length (bp)</b>	561,095,905	562,329,395	552,971,862	554,538,566	542,077,279	543,922,345
<b>Genome repeat length (bp)</b>	242,454,194	242,987,195	247,826,324	248,528,477	226,598,236	227,369,507
<b>Genome unique length (bp)</b>	318,641,711	319,342,200	305,145,537	306,010,089	315,479,043	316,552,838
<b>Model Fit</b>	63.17%	97.85%	61.64%	97.57%	63.90%	97.63%
<b>Read error rate</b>	0.38%	0.38%	0.43%	0.43%	0.19%	0.19%

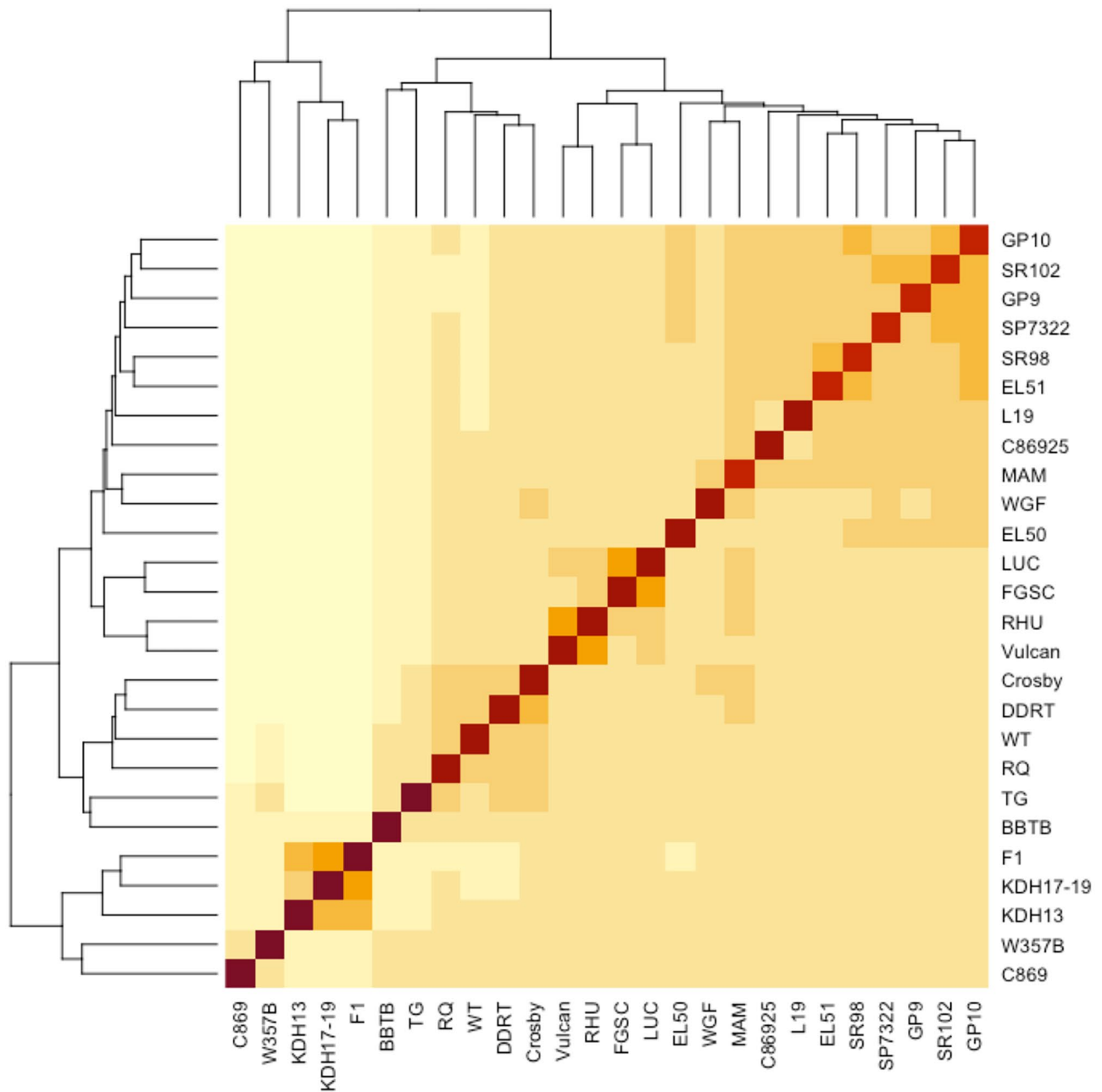
\*  $p=2$ ,  $k=21$

**Table 7** Variation detected within KDH13, KDH19-17, and  $F_1$  (KDH13 × KDH19-17) using reference genome EL10.1

	KDH13		KDH19-17		$F_1$ (KDH13 × KDH19-17)		Total	
Total	1,243,131	100.0%	1,779,921	100.0%	2,526,295	100.0%	3,086,720	100.0%
SNP	958,440	77.1%	1,321,254	74.2%	1,948,743	77.1%	2,259,324	73.2%
Insertions	30,740	2.5%	111,223	6.2%	158,838	6.3%	191,448	6.2%
Deletions	69,295	5.6%	114,347	6.4%	164,200	6.5%	198,057	6.4%
Complex	74,036	6.0%	148,467	8.3%	222,444	8.8%	268,090	8.7%
MNP	82,007	6.6%	50,051	2.8%	75,442	3.0%	90,004	2.9%
SV	28,613	2.3%	34,579	1.9%	16,605	0.7%	79,797	2.6%
Heterozygous	76,388	6.1%	234,262	13.2%	1,775,332	70.3%		

**Fig. 3** Plot of genome variation in KDH13, KDH19-17, and  $F_1$  (KDH13xKDH19-17) relative to the *Beta vulgaris* EL10.1 genome





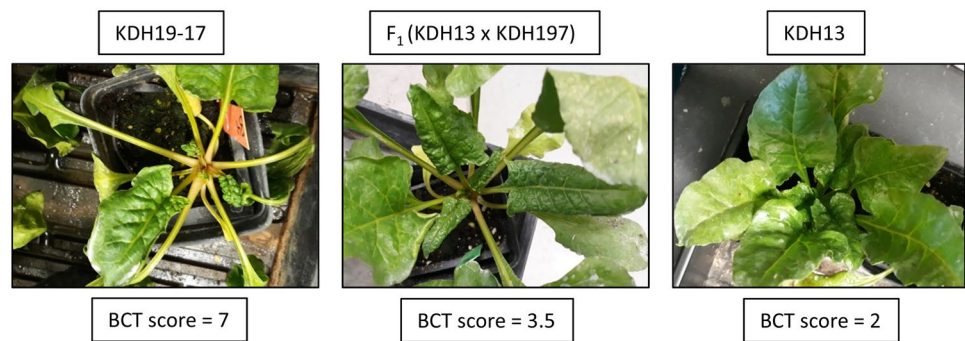
**Fig. 4** Hierarchical clustering based on relationship coefficients determined between KDH13, KDH19-17, and  $F_1$  (KDH13  $\times$  KDH19-17) relative to other sequenced genomes representing the cultivated lineages of *Beta vulgaris*

ancestral lineage. Most obvious is the divergence between the C869 single plant versus C86925, which represents a pool of 25 individuals. KDH13 resolves closer to C869 than KDH19-17; this was also seen in an increase in the number of indels and SNP shared between these two lines, suggesting a greater degree of shared variation. The population C869 represents the population from which EL10.1 was derived. When clustering the data based on relationship coefficients, the  $F_1$  resolved intermediate to the two parents KDH13 and KDH19-17, providing additional support for the hybrid nature of the  $F_1$  genome.

#### Phenotypic Data (KDH13, KDH19-17, and $F_1$ (KDH13 $\times$ KDH19-17))

KDH13 showed a greater degree of resistance to BCTV relative to KDH19-17 (rated 2.0 on a scale of 0 to 9). No variability within these populations suggests a highly reproducible resistance phenotype has been captured within the KDH13 genetic background. KDH19-17 proved to be susceptible (rated 7.0), making it a good parent in order to develop  $F_2$ -derived populations where recombinants of the parental genomes can be

**Fig. 5** Beet curly top disease scores for  $F_1$  (KDH13  $\times$  KDH19-17) and parental lines KDH13 and KDH19-17



evaluated. The  $F_1$  population all showed an intermediate phenotype (rated 3.5) (Fig. 4). Although comparisons of phenotypic data between contrasting parents and the  $F_1$  cannot determine whether alleles have dominant, codominant, and recessive gene action of alleles, the intermediate phenotype of the  $F_1$  indicates the genetic variation as conditioning resistance to BCTV is heritable. Both the phenotypic data and marker data support the hybrid nature of the population and further analysis of  $F_2$ -derived populations in order to determine what genomic variation influences BCTV resistance (Fig. 5).

## Discussion

Long-read PacBio sequence data was generated for the assembly of a genome for BCTV-resistant genetic stock KDH13 (PI 663862). The ability to capture of genome sequence data associated with agronomically important lineages is the backbone of a “genome-informed” approach to crop improvement. The knowledge and resources developed from genome sequencing can make plant breeding, genetic resource conservation, and functional genomics programs more effective and efficient. The identification and sorting of functional variation within genomes associated with crop productivity and sustainability allows the efficient transfer traits into elite cultivars via genetic markers. If traits are understood from a molecular perspective or by association, traits such as durable genetic resistance can be deployed more rapidly to address grower challenges.

Sequence data of KDH13, KDH19-17, and an  $F_1$  individual (KDH19-17  $\times$  KDH13) ( $F_1$ ) allowed for determination of KDH13-specific variation. In total, 1,272,863 variants were informative for distinguishing parental genome segments in recombinant  $F_2$ -derived populations. Methods which look at phenotypic bulks within segregating  $F_2$ -derived populations will likely be able to reduce the number of variants to a more tractable number for breeding. In fact, relatively simple population designs coupled with next-generation sequencing have been successful in identifying causal variation in beet (Ries et al. 2016) using a mapping by sequencing approach (Hartwig et al. 2012). The ability to produce DH lines and

inbred populations can provide a stable source for important characters which can be readily used by seed companies as a source to enhance characters such as disease resistance.

The broad similarity observed between sugar beet genomes RefBeet-1.2.2, EL10.1, and KDH13 was evident. This included a high read alignment percentage, clustering based on relationship coefficients, relative proportions nucleotides, and genome assembly size. More in-depth comparisons regarding comparisons between EL10.1 and RefBeet-1.2.2 are available in McGrath et al. (2020). Sugar beet genomes share a high degree of genome variation and cluster at the population level (Galewski and McGrath 2020). Perhaps not surprising given the high degree of conservation of gene order that was observed within phylogenetically distant beta taxa (del Rio et al. 2019), as well as the known domestication history and putative genetic bottlenecks which served to fix important traits but may limit the potential for future improvement (Hoffmann and Kenter 2018). Using different sequencing technologies may be a way to characterize and manage diversity better (e.g., SNP and indels using Illumina, SV and chromosome inversions using PacBio, and chromosome level variation, using Hi-C, Bio Nano, and linkage maps). Diversity can be deleterious as well as beneficial; it is needed for improvement because selection serves to purge diversity as beneficial traits become fixed. The genetic base of modern corn lineages exhibits a 43% reduction in genetic diversity relative to progenitor populations (Wright et al. 2005). Boarding the genetic base of important agricultural crops including has been proposed in beet in order to gain important agronomic characters such as disease resistance (Panella and Lewellen 2007). This has been identified as an issue in many crops to the degree that “re-domestication” efforts are now being considered (Fernie and Yan 2019).

Long reads have been critical for characterizing structural variants in crop species. PacBio has been used in major crops to address this question by examining chromosome level variation within the species (Sun et al. 2018; Zhang et al. 2016). High-quality reference genomes of beet appear to require additional technologies to achieve chromosome level scaffolding such as Hi-C and Bio nano optical maps. Genetic linkage maps were used to order scaffolds

of the RefBeet-1.2 genome assembly. Evaluation of beet genome assemblies has benefited from knowledge of physical cytogenetic marks discovered using fluorescence in situ hybridization (Paesold et al. 2012). Genome quality was highly variable as a result of different sequencing strategies and technologies. RNAseq was generated for the EL10.1 and RefBeet-1.2.2 which provided additional evidence for gene models. The RefBeet-1.2.2 assembly generated full-length CDS from single-molecule sequencing which may have increased the accuracy of gene annotation and estimates of gene content. RNAseq will continue to be an effective tool to annotate genes present in diverse genomes but not the reference. Pan genome analysis is becoming necessary to determine how genome composition affects phenotypic differences within and between lineages (Tao et al. 2019). In beets, as in many important crops, disease resistance is the hallmark of breeding efforts and critical for sustainability and future production of the species within conventional agricultural systems. Many of these genes are likely highly duplicated and may be located in repetitive regions of the genome (Funk et al. 2018). This makes long-read sequencing more attractive for defining these targets. As new material is sequenced for insight into traits of agronomic importance, we can assume greater levels of chromosome level variation will be present with greater genetic distance. The KDH13 genome provides a resource and insight into variation of high agronomic value which may not be present in current reference genomes. Future work will determine if PacBio sequencing alone can capture the unique genomic variation associated with BCTV resistance and whether the positional information and gene models leveraged from existing reference genomes is sufficient to compare the order and orientation of genome variation as well as functional gene content between diverse beet genomes.

**Author Contribution** PJG wrote the manuscript draft and carried out the computational analysis; IE developed materials for sequencing and provided the PacBio and Illumina sequencing data.

**Funding** United States Department of Agriculture CRIS project #2054–21220-005-00D.

**Availability of Data and Material** KDH13 PacBio reads were deposited to NCBI under BioProject (PRJNA665499), KDH13 Illumina data (BvvSeq-1 PRJNA176558), F<sub>1</sub>, and KDH19-17 (PRJNA665499). The EL10.1 genome assembly (PRJNA413079) and the beet crop type populations (PRJNA563463). Augustus gene predictions, PfamScan results, KDH13 Assembly (Canu 1.7), ordered KDH13 assembly, vcf files and files used for visualization are available at (Data Dryad—<https://datadryad.org/stash>).

**Code Availability** All code is available at [https://github.com/BetaGenomeNinja-/PacBio\\_KDH13](https://github.com/BetaGenomeNinja-/PacBio_KDH13). This includes bash scripts for file manipulation and command line usage of programs which can be sourced via bioconda (<https://anaconda.org/bioconda>). R

code for ordering the genome sequence, evaluating blast results, visualization of genome variation and genome order with respect to the EL10.1 genome is also available.

## Declarations

**Consent to Participate** Both authors consent to participate.

**Consent for Publication** Both authors consent to the publication of this research.

**Competing Interests** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. *Plant Mol Biol Report* 9:208–218
- Bevan MW, Uauy C, WULFF BBH, Zhou J, Krasileva K et al (2017) Genomic innovation for crop improvement. *Nature* 543:346–354
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120
- Bradnam KR, Fass JN, Alexandrov A et al (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2:10
- Butterfass T (1964) Die chloroplastenzahlen in verschiedenartigen zel- len trisomer zuckerruben (*Beta vulgaris* L.). *Z Bot* 52:46–77
- Chawla HS, Lee H, Gabur I et al (2020) Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. *Plant Biotechnol J*. <https://doi.org/10.1111/pbi.13456>
- Chen JQ, Wu Y, Yang H, Bergelson J, Kreitman M et al (2009) Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol Biol Evol* 26:1523–1531
- Chen X et al (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32:1220–1222. <https://doi.org/10.1093/bioinformatics/btv710>
- del Rio AR, Minoche AE, Zwickl NF, Friedrich A, Liedtke S, Schmidt T, Himmelbauer H, Dohm JC (2019) Genomes of the wild beets *Beta patula* and *Beta vulgaris* ssp. *maritima*. *The Plant J*. <https://doi.org/10.1111/tpj.14413>
- Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F et al (2014) The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* 505:546–549
- Eujayl IA, Strausbaugh CA, Lu C (2016) Registration of sugar beet doubled haploid line KDH13 with resistance to beet curly top. *J Plant Reg* 10:93–96
- Fernie AR, Yan J (2019) De novo domestication: an alternative route toward new crops for the future. *Mol Plant* 12:615–631

- Funk A, Galewski P, McGrath JM (2018) Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome. *Plant J* 95:659–671
- Galewski P, McGrath JM (2020) Genetic diversity among cultivated beets (*Beta vulgaris*) assessed via population-based whole genome sequences. *BMC Genomics* 21:189
- Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint*. <https://arxiv.org/abs/1207.3907>
- Hartwig B, James GV, Konrad K, Schneeberger K, Turck F (2012) Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiol* 160(2):591–600
- Hoffmann CM, Kenter C (2018) Yield potential of sugar beet – have we hit the ceiling? *Front Plant Sci* 9:289
- Jayakumar V, Sakakibara Y (2019) Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief Bioinform* 20(3):866–876
- Kowar T, Zakrzewski F, Macas J, Kobližková A, Viehovec P, Weisshaar B, Schmidt T (2016) Repeat composition of CenH3-chromatin and H3K9me2-marked heterochromatin in Sugar Beet (*Beta vulgaris*). *BMC Plant Biol* 16(1):120. <https://doi.org/10.1186/s12870-016-0805-5>
- Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. <https://arxiv.org/abs/1303.3997v1>
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M et al (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26:2867–2873
- Marçais G, Kingsford CA (2011) fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6). <https://doi.org/10.1093/bioinformatics/btr011>
- McGrath JM, Funk A, Galewski P, Ou S, Townsend B et al (2020) A contiguous de novo genome assembly of sugar beet EL10 *Beta vulgaris* L. *bioRxiv*.298315
- Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A (2018) Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 34(13):142–150
- Mumford DL (1974) Procedure for inducing curly top epidemics in field plots. *J Am Soc Sugar Beet Tech* 18:20–23
- Ou S, Jiang N (2018) LTR\_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* 176:1410–1422
- Paesold S, Borchardt D, Schmidt T, Dechyeva D (2012) A sugar beet (*Beta vulgaris* L.) reference FISH karyotype for chromosome and chromosome-arm identification, integration of genetic linkage groups and analysis of major repeat family distribution. *Plant J* 72:600–611
- Panella L, Lewellen RT (2007) Broadening the genetic base of sugar beet: Introgression from wild relatives. *Euphytica* 154:383–400
- Panella L, Kaffka SK, Lewellen, RT, McGrath JM, Metzger MS, Strausbaugh CA (2014) Sugarbeet. Pages 357–396 in: *Yield Gains in Major U.S. Field Crops*. Crop Sci. Soc. Am. Spec. Publ. 33. Smith S, Diers B, Specht J, and Carver B (eds). Crop Science Society of America, Madison, WI
- Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11:1432. <https://doi.org/10.1038/s41467-020-14998-3>
- Ries D, Holtgräwe D, Viehöver P, Weisshaar B (2016) Rapid gene identification in sugar beet using deep sequencing of DNA from phenotypic pools selected from breeding panels. *BMC Genomics* 17:236
- Schiessl SV, Kathe E, Ihlen E, Chawla HS, Mason SA (2019) The role of genomic structural variation in the genetic improvement of polyploid crops. *Crop J* 7:127–140
- Schondelmaier J, Jung C (1997) Chromosomal assignment of the nine linkage groups of sugar beet (*Beta vulgaris* L.) using primary trisomics. *Theor Appl Genet* 95:590–596
- Seppely M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M. (eds) *Gene Prediction. Methods in Molecular Biology*, vol 1962. Humana, New York, NY. 2019. [https://doi.org/10.1007/978-1-4939-9173-0\\_14](https://doi.org/10.1007/978-1-4939-9173-0_14)
- Shumate A, Salzberg SL (2020) Liftoff: accurate mapping of gene annotations. *Bioinformatics*, December. <https://doi.org/10.1093/bioinformatics/btaa1016>
- Strausbaugh CA, Gillen AM, Gallian JJ, Camp S, Stander JR (2006) Influence of host resistance and insecticide seed treatments on curly top in sugar beets. *Plant Dis* 90:1539–1544
- Strausbaugh CA, Wenninger EJ, Eujayl IA (2012) Management of severe curly top in sugar beet with insecticides. *Plant Dis* 96:1159–1164
- Strausbaugh CA, Wenninger EJ, Eujayl IA (2014) Control of curly top in sugar beet with seed and foliar insecticides. *Plant Dis* 98:1075–1080
- Sun et al (2018) Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat Genet* 50:1289–1295
- Tao YF et al (2019) Exploring and exploiting pan-genomics for crop improvement. *Mol Plant* 12:156–169
- Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF et al (2005) The effects of artificial selection on the maize genome. *Science* 80(308):1310–1314
- Zhang JW et al (2016) Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc Natl Acad Sci* 113:5163–5171

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.